

RESEARCH

Open Access



# Building a glaucoma interaction network using a text mining approach

Maha Soliman<sup>1\*</sup>, Olfa Nasraoui<sup>2</sup> and Nigel G. F. Cooper<sup>1</sup>

\* Correspondence:

maha.soliman@louisville.edu

<sup>1</sup>Department of Anatomical Sciences and Neurobiology, University of Louisville, School of Medicine, Louisville, KY, USA  
Full list of author information is available at the end of the article

## Abstract

**Background:** The volume of biomedical literature and its underlying knowledge base is rapidly expanding, making it beyond the ability of a single human being to read through all the literature. Several automated methods have been developed to help make sense of this dilemma. The present study reports on the results of a text mining approach to extract gene interactions from the data warehouse of published experimental results which are then used to benchmark an interaction network associated with glaucoma. To the best of our knowledge, there is, as yet, no glaucoma interaction network derived solely from text mining approaches. The presence of such a network could provide a useful summative knowledge base to complement other forms of clinical information related to this disease.

**Results:** A glaucoma corpus was constructed from PubMed Central and a text mining approach was applied to extract genes and their relations from this corpus. The extracted relations between genes were checked using reference interaction databases and classified generally as known or new relations. The extracted genes and relations were then used to construct a glaucoma interaction network. Analysis of the resulting network indicated that it bears the characteristics of a small world interaction network. Our analysis showed the presence of seven glaucoma linked genes that defined the network modularity. A web-based system for browsing and visualizing the extracted glaucoma related interaction networks is made available at <http://neurogene.spd.louisville.edu/GlaucomaINViewer/Form1.aspx>.

**Conclusions:** This study has reported the first version of a glaucoma interaction network using a text mining approach. The power of such an approach is in its ability to cover a wide range of glaucoma related studies published over many years. Hence, a bigger picture of the disease can be established. To the best of our knowledge, this is the first glaucoma interaction network to summarize the known literature. The major findings were a set of relations that could not be found in existing interaction databases and that were found to be new, in addition to a smaller subnetwork consisting of interconnected clusters of seven glaucoma genes. Future improvements can be applied towards obtaining a better version of this network.

**Keywords:** Text mining, Interaction network, Glaucoma, Relation extraction

## Background

Extraction of biological networks, related to specific diseases or conditions from the scientific literature, is an emerging problem which may be solved with the aid of text mining approaches. Biological networks are important features used for modelling, analysis and simulation of biological systems [1], and for the development of hypotheses from data-

sets [2–6]. In general, the inference of an interaction network from text can be sub-tasked as: 1) determination of the source of the text to be searched, 2) identification of the entities to be extracted (genes, proteins, metabolites, diseases), and 3) inference of potential relationships between selected entities. Once these subtasks are resolved, the entities and their relationships can be mapped to the nodes and edges of a biological network. A common aspect for subtasks two and three is their amenability to the use of text mining methods for their resolution.

As for the first subtask, the source of text to be mined can be abstracts or full text articles in collections of scientific publications. While the use of abstracts would be more advantageous due to their concise information content [7–9], an increasing number of text mining approaches make use of full text journals [10]. However, in trying to deal with full text publications, there are technical challenges due to the existence of different formats (pdf, HTML) as well as non-uniform substructure across journals. In terms of the second subtask, there are many examples in the literature in which text mining approaches have been used to infer a relationship between biomarker genes and diseases/disorders, including for example, insulin-resistance [11], Alzheimer disease [12], breast cancer [13], prostate cancer [14], and respiratory disease [15]. Therefore, it is possible to develop putative associations between biomarkers and glaucoma with a text mining approach. The third sub-task is to develop a relation extraction (RE) process to reliably infer binary relationships between the entities previously derived from subtask one. Relationships depend on the type of entities we are dealing with. For example, if an entity is a transcription factor, then the textual terms that reflect regulation (up/down-regulate..., etc.) can be sought in the relation extraction process. If an entity is a protein, then textual terms that reflect activation or binding are sought in the relation extraction process [16, 17]. RE can be a closed or an open process. It is closed when there is a set of relations determined a priori such as, (“activate”, “up-regulate”, “express”) and the extractor predicts one of a finite and fixed set of relations. It is open when no relations are specified in advance [18]. For example, an open RE system that runs over the sentence “*HSPA6* is a potential target gene of *FOXCI*”, will list the following binary relation:

(*HSPA6*, is a target gene, *FOXCI*)

On the other hand, if a closed RE is used, this relation will not be extracted unless the relation “target” was included in the set of relations determined a priori. In general, a closed RE is useful when extracting relations from scientific literature, while an open RE is suitable when extracting relations from the web [19].

Text mining services have evolved rapidly to become an important component of inference pipelines. The next generation of text mining approaches have to deal with the construction of complete text mining systems to aid the inference of interactions or associations between bio entities. OntoGene [20], Anni [21], RLISM [22], and CRAB [23] are examples of such next generation systems. In terms of usage, OntoGene is considered the most integrative because it allows the detection of entities and relationships from selected categories of entities, such as proteins, genes, drugs, diseases, and chemicals. On the other hand, Anni has the advantage of introducing an ontology based interface to MEDLINE, and it is capable of retrieving documents for several classes of biomedical concepts. In addition, RLIMS-P and CRAB 2.0 are topic specific approaches.

For example, RLIMS-P targets protein phosphorylation and CRAB 2.0 targets cancer risk assessment.

The goal of this study is to initiate the development of a glaucoma interaction network with the aid of text mining the open access scientific literature housed in PubMed Central (PMC). According to the Glaucoma Research Foundation (GRF), glaucoma is the second leading cause of blindness [24]. It is an invisible disease and gradually steals sight without warning. Generally, it cannot be cured, but it can be controlled [25]. Some reported glaucoma interaction networks were based on genome wide association studies (GWAS) [26, 27] while others focused on interaction networks from genome wide expression studies (GWES) [28, 29] but none have yet been based solely on text mining of the vast swath of PMC literature, where all types of glaucoma studies are covered. Such a network is expected to have a wider coverage than prior efforts because it will not be inferred from a particular type of study but rather from all types of studies related to glaucoma.

## Methods

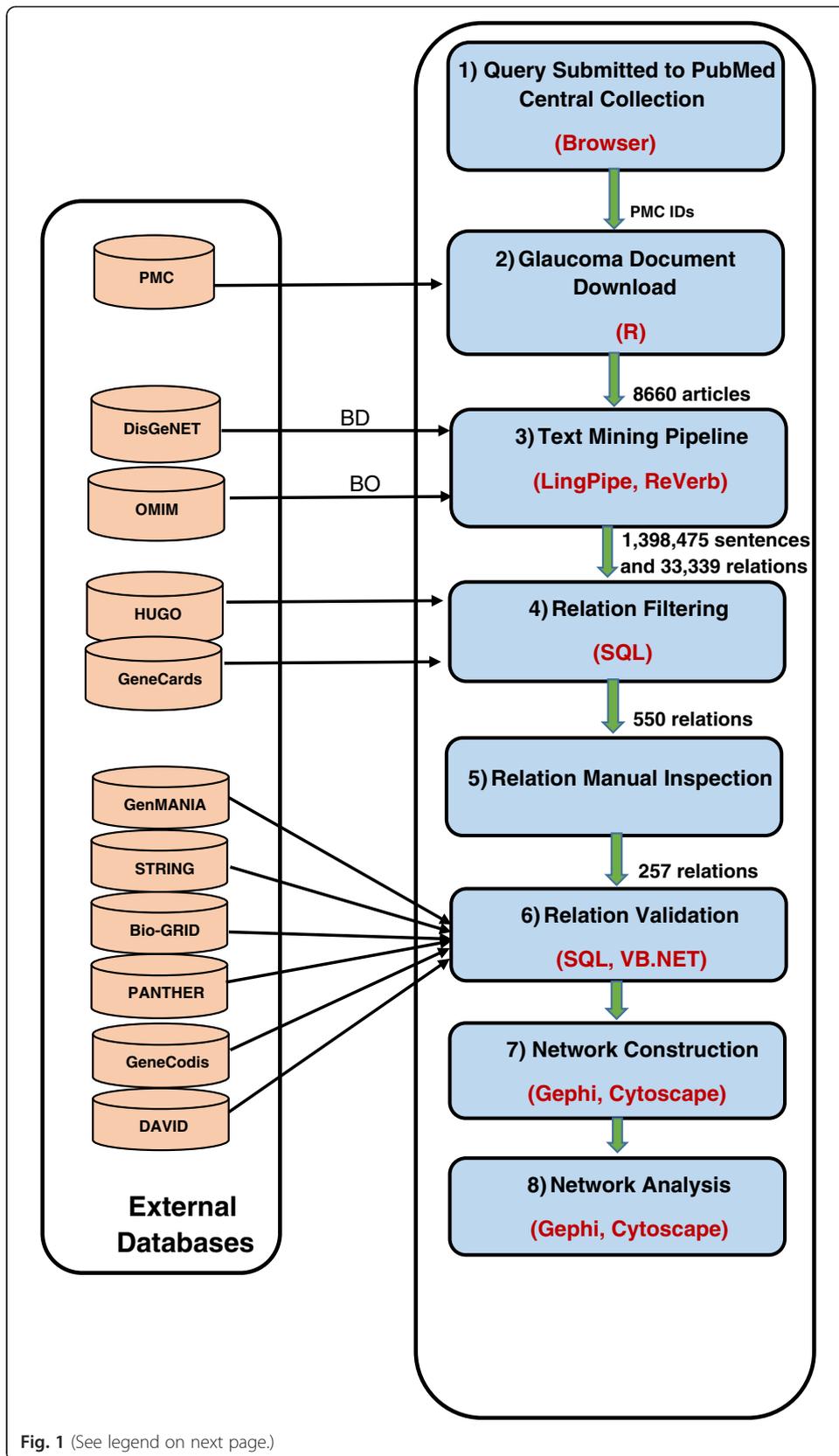
Text mining enables the discovery of useful knowledge from unstructured or semi-structured text [30, 31] which fits the goal of this study. Figure 1 is the flow diagram that shows how the results in this study are generated. The text mining pipeline (Fig. 2), which was used in step 3 of the flow diagram, starts from each article containing some information to be extracted. The article is first segmented into its constituent sentences using a segmenter. Each sentence is then sub-segmented into its constituent words, called tokens, using a tokenizer. Subsequently, part of speech (POS) tagging is applied to each of the tokens to identify the role of each word within the sentence. Additionally, a name entity recognition (NER) is used to identify target entities, which are gene names. Finally, a relation extraction (RE) routine is applied to extract existing relations within each sentence. The relations are then validated, where possible, against an existing reference knowledgebase. Finally, entities and relations are translated into an interaction network. The main tasks in our methodology are:

### Text selection and retrieval

Unlike PubMed, all articles in PubMed Central (PMC) are full text and open access. This makes PMC a suitable repository of the literature for mining full text articles. We used a PubMed medical subject headings (MeSH) terms query to collect all possible glaucoma related articles. PMC Open Access was queried for eight types of key terms related to glaucoma including: “open-angle glaucoma”, “angle-closure glaucoma”, “secondary glaucoma”, “congenital glaucoma”, “hyper glaucoma”, “neovascular glaucoma”, “pigmentary dispersion glaucoma” and “open access”. The resulting data set composed a corpus of 8,660 full length articles ready for mining. Articles were downloaded from PMC Open Access according to the PMC OAI service [32].

### Entity selection and extraction

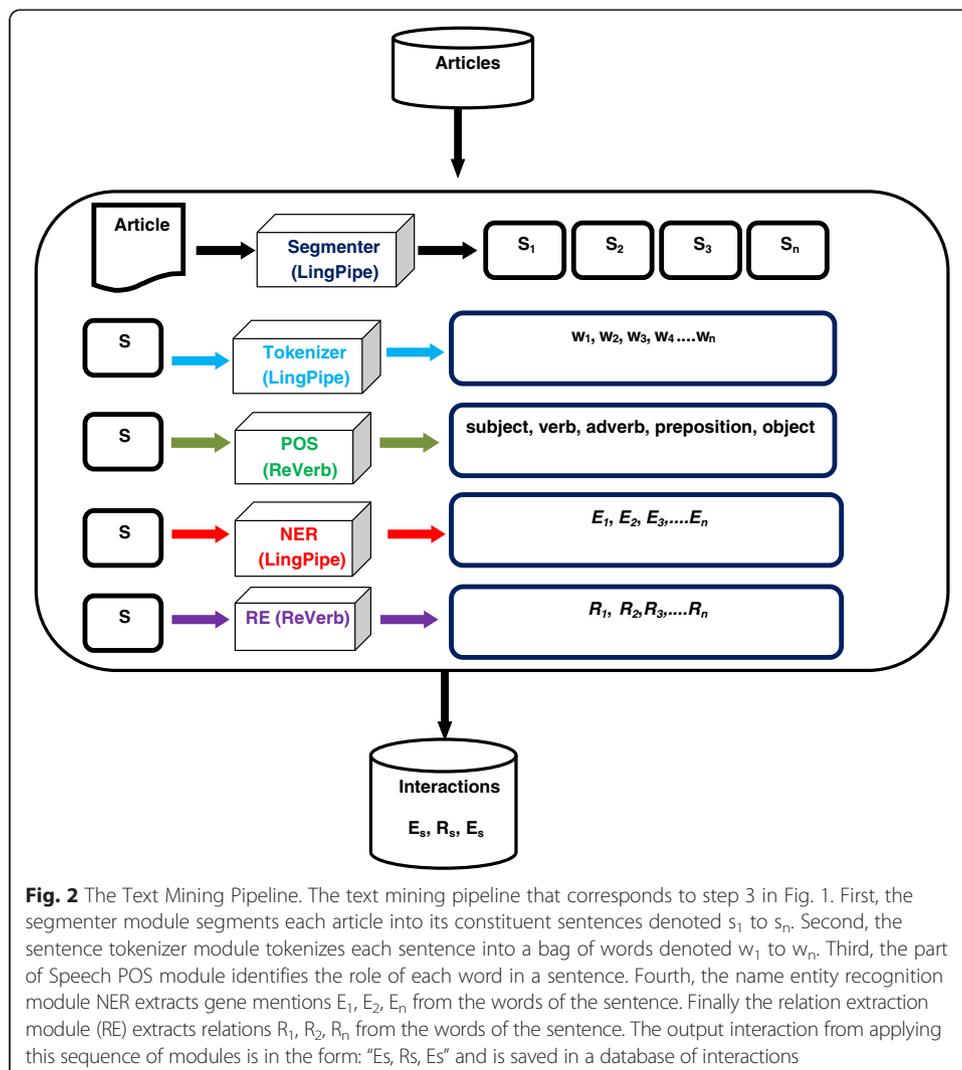
This study targets the extraction of gene associations which have been previously linked to glaucoma in the open access literature. Our target entities, broadly speaking, are “gene/gene products”. In our approach, we did not make any distinction between



(See figure on previous page.)

**Fig. 1** The workflow pipeline followed to build the glaucoma interaction network. Step 1: PubMed Central is queried for glaucoma related articles. Step 2: all glaucoma articles are collected and a glaucoma collection is constructed. Step 3: each document in the resulting collection is processed using the text mining pipeline detailed in Fig. 2 and a set of relations is obtained. Step 4: relations are stored into a database and filtered using SQL queries. Step 5: Filtered relations are subjected to manual inspection to identify meaningful relations worthy of validation. Step 6: inspected relations are then validated and evaluated against external reference databases. Step 7: validated relations are mapped to nodes and edges to form a potential glaucoma network. Step 8: network analysis of the resulting network is performed. The left panel contains external databases needed by each step of the workflow. See Table 1 for definition of BD, and BO

mentions of gene, mRNA, or protein in the text. For simplicity, we will reference gene/gene products as “gene”. Association can cover direct protein-protein interaction (PPI) type; predicted or found experimentally, bimolecular events such as expression and localization, and/or static relations. Our definition for association is a loose biological definition that covers any relation that holds between genes or related entities, that is of biological/biomedical or health-related interest, without necessarily implying change [33, 42]. It is for this reason that we have opted for an open RE strategy.



Our glaucoma corpus was segmented into 1,398,475 sentences with the LingPipe sentence segmenter [34]. Genes within sentences were annotated using the LingPipe taggers CharLmHmmChunker and TokenShapeChunker. The performance of any tagger can be evaluated by testing the tagger on an annotated corpus. GenTag [35], and GENIA [36] are well known biomedical annotated corpuses for performance evaluation of taggers. CharLmHmmChunker is trained on GenTag while TokenShapeChunker is trained on GENIA. Compared to GENIA, GenTag is more generic and less specific while GENIA has annotations for 36 biomedical named entities, and therefore provides a breadth classification. Our motivation for using both taggers is to maximize the number of extracted genes [37]. Both taggers accept full length articles as text files and provide an output of annotated files, formatted in Standard Generalized Mark-up Language (SGML) for gene mentions. SGML uses XML tags to describe a mentioned gene but the user will need to specify an encoding system for both input and output files, as well as the desired type of input/output files. For our particular study, we have used the "UTF-8" encoding system, and plain text format for our input/output files.

### Benchmarking genes

A total of 305 glaucoma benchmark genes (BG) were used in this study. Of this number, 155 come from the Online Mendelian Inheritance in the Man database, OMIM® [38] (BO), while the 180 remaining genes come from the Disease Gene Network database DisGeNET release 2.1.0 (July 2014) [39–41] (BD). There were 30 benchmark genes (BC) common to both OMIM and DisGeNET databases (Table 1) indicating their likely importance to glaucoma. The union of OMIM and DisGeNET genes were used as benchmark genes for our intended glaucoma interaction network (Additional File 3). Table 1 lists the benchmark gene types and their abbreviations. Any gene in the literature, which was co-listed in one sentence with one of these BG, is considered a putative association. Sentences, that contain one gene, were filtered out from the tagged sentences to focus our search on sentences that have two or more genes, provided that one of the genes was a BG. If the sentence does not contain a BG, then it is excluded. The idea of the filtering step was to ensure the existence of interacting genes with some BG. The next task is to capture associations between the BG and other non-benchmark genes (NBG), thus constructing a glaucoma interaction network capturing potentially novel relations. The

**Table 1** Glaucoma benchmark and non-benchmark genes used in building the network

Abbreviation	Definition	Number	Percent
BO	Benchmark glaucoma genes from OMIM database queried with "Glaucoma"	155	51 %
BD	Benchmark glaucoma genes from DisGeNET database queried with "Glaucoma"	180	59 %
BC	Benchmark glaucoma genes from the intersection of OMIM and DisGeNET databases	30 (BO∩BD)	10 %
BG	Benchmark glaucoma genes from union of BO and BD	305 (BO∪BD)	100 %
NBG	Non-benchmark genes from PubMed Central	150	N/A

For simplicity, benchmark genes used to build the interaction network are abbreviated as BG. If BG are obtained from OMIM, then we call them BO. If BG are obtained from DisGeNET, then we call them BD. Benchmark genes, common to OMIM and DisGeNET, are called BC. Genes that are not benchmark genes are called NBG. The definition, number and percentages of all benchmark genes are listed in columns 2 to 4

output of this step is a list of associated genes. Some genes were found to be a gene name, a gene synonym, or a previous gene symbol and all of these aliases were mapped to their HUGO approved gene symbol [42].

### Relation extraction

Sentences that contain putative pairs were subjected to the open source relation extractor ReVerb [43] to extract binary relationships between gene mentions. ReVerb parses each sentence and identifies its main verb. It then starts identifying the subject and object of the sentence. It outputs triplets of “E, Rel, E”, where E is an entity and Rel is a relationship (the main verb of the sentence). In addition to extracted relations, ReVerb also outputs a confidence score associated with the relation that reflects how much ReVerb is certain of its extraction mechanism. Application of ReVerb identified 33,339 binary relations. Extracted relations were verified using the interaction databases GeneMANIA [44] and the Biological General Repository for Interaction Datasets database (BioGRID release 3.4.129) [45]. If the reference databases could not recognize a particular gene in a relation, the gene’s different aliases are first retrieved from GeneCards [46] and the relation is verified using GeneMANIA or BioGRID.

### Network construction

Extracted entities and relations were manually inspected and mapped to nodes and edges. The Gephi open source graph visualization software tool [47] was used to develop a graphic representation of the extracted interaction network (Fig. 7). Analysis of the generated network was carried out with the Cytoscape network analyzer [48]. Enrichment analysis for the extracted genes was conducted through the PANTHER classification system version 10.0 (release May 2015) [49], as well as the Database for Annotation, Visualization and Integrated Discovery (DAVID) [50, 51], and the gene annotations co-occurrence discovery database (GeneCodis) [52–54].

### Results

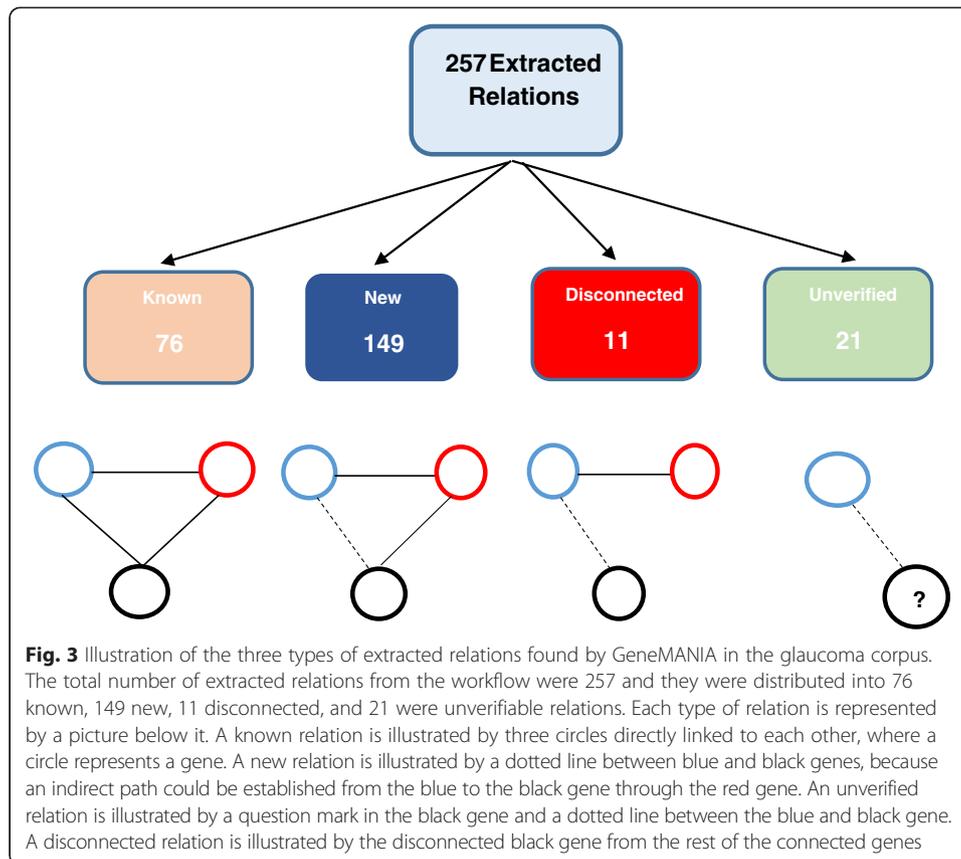
The output from ReVerb may contain incorrect triplets. Therefore, all triplets were saved into a database and were subjected to a filtering process, in which a query is constructed to extract triplets that contained any biological entity name. Filtering ReVerb relations resulted in a total of 550 triplets of “E, Rel, E”, where E is an entity (gene), and Rel is a verb associating the two entities. Some relations from the filtered list of the 550 relations involved “POAG” (Primary Open Angle Glaucoma), while others involve “XFS” (Exfoliation syndrome), a developmental variant of glaucoma (Table 2). The relations included known relations, new relations, disconnected relations, redundant relations, misinterpreted relations, and unverified relations. A known relation is a previously published relation, for example, the relation between *OPTN* and *MYOC*. A relation is defined as new when no direct link between its entities is reported by GeneMANIA or BioGRID. If an indirect link can be established between relation entities through an intervening gene(s), then it is evidence for the possibility of the relation. If no indirect link can be established between relation entities, then it is a disconnected relation, in other words, a relation involving nodes that are currently considered to be disconnected. A redundant relation is a known or new relation, but is repeated many times. A misinterpreted relation is a relation involving an

**Table 2** Genes related to Primary Open Angle Glaucoma (POAG) and Exfoliation syndrome (XFS)

Gene	Disease	Confidence	Support
<i>MYOC</i>	POAG	0.98	30
<i>LOXL1</i>	XFS	0.98	12
<i>TG</i>	XFS	0.98	1
<i>CYP1B1</i>	POAG	0.97	12
<i>GSTT1</i>	POAG	0.97	4
<i>CAV1</i>	POAG	0.97	2
<i>SPARC</i>	POAG	0.96	2
<i>CPE</i>	POAG	0.96	1
<i>APOE</i>	POAG	0.94	7
<i>CDKN2B-AS1</i>	POAG	0.94	3
<i>OPTN</i>	POAG	0.93	17
<i>NOS3</i>	POAG	0.93	5
<i>WDR36</i>	POAG	0.92	13
<i>GLC1A</i>	POAG	0.92	2
<i>GLC1N</i>	POAG	0.92	2
<i>GSTM1</i>	POAG	0.91	4
<i>PDI45</i>	POAG	0.89	2
<i>GC</i>	XFS	0.88	1
<i>T</i>	XFS	0.88	2
<i>TTR</i>	POAG	0.87	2
<i>LOXL1</i>	POAG	0.86	3
<i>CDKN2B</i>	POAG	0.86	2
<i>SIX1</i>	POAG	0.85	2
<i>NTF4</i>	POAG	0.83	4
<i>CNTNAP2</i>	XFS	0.83	1
<i>GLC3A</i>	POAG	0.82	2
<i>OPA1</i>	POAG	0.81	2
<i>TBK1</i>	POAG	0.78	2
<i>MMP1</i>	XFS	0.67	1
<i>MMP3</i>	XFS	0.67	1
<i>TP53</i>	POAG	0.66	1
<b><i>ELN</i></b>	<b>XFS</b>	<b>0.24</b>	<b>1</b>

The gene and its related disease are listed under the "Gene" and "Disease" columns respectively. The confidence column is the maximum of all confidence values reported by ReVerb for the same relation, extracted from multiple articles. Relations with low confidence are bolded. The support column is the count of articles listing the same gene relation

acronym that is identical to a gene symbol, for example *ECD* is an acronym for the endothelial cell density, but was captured as a gene symbol for ecdysoneless homolog gene. An unverified relation is a known or new relation, involving a gene that is not identified by HUGO, GeneCards, or GeneMANIA. Filtering out redundant and misinterpreted relations resulted in a total of 257 unique triplets (E REL E), that include 74 genes from the combined DisGeNet and OMIM databases (BG), 17 of which were common (BC) to both databases (BO, BD), and 150 related genes (NBG) uncovered from the PubMed Central literature database. In terms of the classification of the extracted relations (Fig. 3), 76 were previously known relations, 149 were new relations, 21 were unverified and yet interpretable relations (Table 3) and 11 relations involved disconnected



nodes, which linkage could not be confirmed at this time (Table 4) and yet some contextual evidence (Column 5 in Table 4) may suggest some plausible linkage. Both of the 550 and the 257 relations can be found in the Additional files 1 and 2 respectively.

### Analysis and validation

The associations between the pair of entities within the 257 extracted triplets (E,Rel,E) were validated against both the GenMANIA database and BioGRID. Validation using BioGRID showed an agreement in only 24 previously known relations with GeneMANIA. Unlike GeneMANIA, BioGRID does not consider the entire gene network for a pair of genes to identify indirect relations as in GeneMANIA. Therefore, all relations, except the 24 known ones, are new according to BioGRID. Most of the 21 unverified relations were due to unrecognized entity symbols in GeneMANIA at the time of writing this paper, such as antisense of a gene (*BDNF-AS*, *CDKN2B-AS*) or small interfering RNA for a particular gene (*siPITX2*, *siCSTA*), microRNA, general protein family name (M-opsin), and gene variants or mutation (*OPTN* variants: *Glu50Lys* or *E50K*). However contextual evidence (text) from PMC-ID papers (col. 7 in Table 4) suggests some evidence based on the experiments reported in the mined literature. A summary of the different extracted relations and their percentages is listed in Table 5 and the top fifty most frequent relations are depicted in Fig. 4.

As mentioned in the results section, the results included 150 NBG in relation with the 74 BG. The 150 NBG were subjected to enrichment analysis through the PANTHER,

**Table 3** Twenty one extracted relations with unverified links from GeneMANIA

Gene1	Gene2	Confidence	Unverified node	PMC Excerpt	PMCID/Year	Remark
<i>CDKN2B-AS1</i>	<i>CDKN2B</i>	0.93	<i>CDKN2B-AS1</i>	CDKN2B-AS1 has been shown to be involved in the regulation of CDKN2B, CDKN2A and ARF expression.	PMC4132588/2014	<i>CDKN2B-AS1</i> is a <i>CDKN2B</i> antisense. GeneMANIA does not recognize gene anti-sense.
<i>CDKN2B-AS1</i>	<i>CDKN2A</i>	0.93	<i>CDKN2B-AS1</i>	CDKN2B-AS1 has been shown to be involved in the regulation of CDKN2B, CDKN2A and ARF expression.	PMC4132588/2014	CDKN2B-AS1 is CDKN2B antisense. GeneMANIA does not recognize gene anti-sense
<i>CDKN2B-AS1</i>	<i>ARF</i>	0.93	<i>CDKN2B-AS1</i>	CDKN2B-AS1 has been shown to be involved in the regulation of CDKN2B, CDKN2A and ARF expression.	PMC4132588/2014	CDKN2B-AS1 is CDKN2B antisense. GeneMANIA does not recognize gene anti-sense
<i>CDKN2BAS</i>	<i>CDKN2A</i>	0.92	<i>CDKN2BAS</i>	CDKN2BAS also regulates the expression of CDKN2A, a gene previously shown to be down-regulated in other neurodegenerative disorders, including Alzheimer's disease, suggesting that regulation of CDKN2A expression by CDKN2BAS could also contribute to degeneration of the optic nerve in glaucoma.	PMC3343074/2012	CDKN2BAS is CDKN2B antisense. GeneMANIA does not recognize gene anti-sense
<i>CNTF</i>	<i>LIFR<math>\beta</math></i>	0.90	<i>LIFR<math>\beta</math></i>	In mouse, human OSM activates the heterodimer of LIF receptor $\beta$ (LIFR $\beta$ and gp130, like CNTF.	PMC4171539/2014	LIFR $\beta$ is a mouse gene that GeneMANIA did not recognize
<i>miR410</i>	<i>VEGFA</i>	0.9	<i>miR410</i>	Protein levels of VEGFA were also down-regulated with miR410 overexpression and up-regulated with miR-410 interference.	PMC400246/2014	GeneMANIA does not recognize microRNAs.
<i>STAT1</i>	<i>ANRIL</i>	0.89	<i>STAT1</i>	The binding of STAT1 induces the expression of ANRIL, and represses CDKN2B in endothelial cells.	PMC3565320/2013	GeneMANIA does not recognize locus <i>ANRIL</i>
<i>siPITX2</i>	<i>DKK1</i>	0.83	<i>siPITX2</i>	DKK1 and KCNJ2 which were shown to be affected by PITX2 siRNAs by real time PCR experiments were each previously reported in one study.	PMC2654047/2009	<i>siPITX2</i> is short interfering <i>PITX2</i> . GeneMANIA does not recognize short interfering RNAs.
<i>siPITX2</i>	<i>KCNJ2</i>	0.83	<i>siPITX2</i>	DKK1 and KCNJ2 which were shown to be affected by PITX2 siRNAs by real time PCR experiments were each previously reported in one study.	PMC2654047/2009	<i>siPITX2</i> is short interfering <i>PITX2</i> . GeneMANIA does not recognize short interfering RNAs.
<i>XCPE1</i>	<i>LTBP2</i>	0.82	<i>XCPE1</i>	LTBP2 was predicted to be regulated by KLF4 (at 10 promoters), SP1 (at eight promoters), GATA4 and TEAD	PMC4019825/2014	<i>XCPE1</i> is X gene core promoter element 1 (DNA element). GeneMANIA does not recognize <i>XCPE1</i>

**Table 3** Twenty one extracted relations with unverified links from GeneMANIA (Continued)

				(at five promoters) and XCPE1 (at four promoters) was associated with LTBP2.		
<i>GLC3A</i>	<i>GLC3B</i>	0.78	<i>GLC3B</i>	To narrow down the potential candidate CNVs (genes) and match the identified CNVs to target regions and/or genes, we first focused on known chromosomal loci for PCG, namely <i>GLC3A</i> (2p2-p21), which harbors <i>CYP1B1</i> , <i>GLC3B</i> (1p36.2-p36.1), and <i>GLC3C</i> (14q23).	PMC3250374/2011	GeneMANIA does not recognize gene locus
<i>GLC3A</i>	<i>GLC3C</i>	0.78	<i>GLC3C</i>	To narrow down the potential candidate CNVs (genes) and match the identified CNVs to target regions and/or genes, we first focused on known chromosomal loci for PCG, namely <i>GLC3A</i> (2p2-p21), which harbors <i>CYP1B1</i> , <i>GLC3B</i> (1p36.2-p36.1), and <i>GLC3C</i> (14q23).	PMC3250374/2011	GeneMANIA does not recognize gene locus
<i>E50K</i>	<i>TBK1</i>	0.74	<i>E50K</i>	Recently, it was found that <i>E50K</i> mutant strongly interacted with <i>TBK1</i> , which evoked intracellular insolubility of <i>OPTN</i> , leading to improper <i>OPTN</i> transition from the endoplasmic reticulum to the Golgi body.	PMC4077773/2014	GeneMANIA recognizes <i>OPTN</i> not its mutated form. <i>E50K</i> is a mutation in the <i>OPTN</i> gene
<i>DCDC4</i>	<i>PAX6</i>	0.74	<i>DCDC4</i>	The 3' deletion identified in family 86 contained <i>ELP4</i> and <i>DCD4</i> , which are located downstream of <i>PAX6</i> .	PMC3044699/2011	<i>DCD4</i> (double cortin domain containing 4) is not found in HUGO
<i>MTMR2</i>	<i>NEFL</i>	0.60	<i>NEFL</i>	However, catalytically inactive CMT disease-related <i>MTMR2</i> mutants lead to <i>NEFL</i> assembly defects and to pathologies similar to the one caused by <i>NEFL</i> mutations, suggesting that <i>MTMR2</i> and <i>NEFL</i> may function in a common pathway in the development and maintenance of peripheral axons.	PMC3514635/2012	GeneMANIA does not recognize <i>NEFL</i> .
<i>TTRV30M</i>	<i>EPO</i>	0.50	<i>TTRV30M</i>	It has been suggested that inhibition of <i>EPO</i> production could be caused by the toxicity of prefibrillar aggregates of <i>TTR V30M</i> .	PMC4087117/2014	GeneMANIA recognizes <i>TTR</i> not its mutated form <i>V30M</i> . <i>V30M</i> is a point mutation within <i>TTR</i>
<b><i>BDNF-AS</i></b>	<b><i>EZH2</i></b>	<b>0.40</b>	<b><i>BDNF-AS</i></b>	Further characterization of <i>BDNF-AS</i> indicates that <i>BDNF-AS</i> recruits <i>EZH2</i> and the <i>PRC2</i> complex to the <i>BDNF</i> promoter to repress <i>BDNF</i> transcription through H3K27me3 histone modifications.	PMC4047558/2014	<i>BDNF-AS</i> is <i>BDNF</i> antisense. GeneMANIA does not recognize anti-sense

**Table 3** Twenty one extracted relations with unverified links from GeneMANIA (Continued)

<b>BDNF-AS</b>	<b>PRC2</b>	<b>0.40</b>	<i>BDNF-AS</i>	Further characterization of BDNF-AS indicates that BDNF-AS recruits EZH2 and the PRC2 complex to the BDNF promoter to repress BDNF transcription through H3K27me3 histone modifications.	PMC4047558/2014	<i>BDNF-AS</i> is <i>BDNF</i> antisense. GeneMANIA does not recognize anti-sense
<b>BDNF-AS</b>	<b>BDNF</b>	<b>0.40</b>	<i>BDNF-AS</i>	Further characterization of BDNF-AS indicates that BDNF-AS recruits EZH2 and the PRC2 complex to the BDNF promoter to repress BDNF transcription through H3K27me3 histone modifications.	PMC4047558/2014	<i>BDNF-AS</i> is <i>BDNF</i> antisense. GeneMANIA does not recognize anti-sense
<b>siCSTA</b>	<b>MYOC</b>	<b>0.35</b>	<i>siCSTA</i>	It would be interesting to investigate whether the application of an inhibitor to CSTA, such as its siRNA, could restore the normal MYOC processing and affect the outcome of the disease.	PMC3352898/2012	<i>siCSTA</i> is short interfering <i>CSTA</i> . GeneMANIA does not cover short interfering RNAs.
<b>Glu50Lys</b>	<b>OPTN</b>	<b>0.3</b>	<i>Glu50Lys</i>	More, recently, Minegishi and coworkers reported that the over-expression of a glaucoma causing-mutation in OPTN, Glu50Lys, produces an accumulation of insoluble OPTN protein that can be blocked with chemical inhibition of TBK1 activity in HEK293 cells.	PMC4038935/2014	<i>Glu50Lys</i> is a mutation in the <i>OPTN</i> gene

The genes in each extracted relation are listed under the "Gene1" and the "Gene2" columns respectively. A measure of confidence, reported by ReVerb, is listed under the "Confidence" column, and relations with low confidence (<0.5) are bolded. The unverified node is listed under the "Unverified node" column. The associated text that relates the two genes is listed under the "PMC Excerpt" column. Some genes were identified by their synonyms found in either GeneCards or GeneMANIA. The PMCID of the original article coupled with the year of publication is given under "PMCID/Year" column. Important remarks and gene synonyms may be listed under the "Remark" column

**Table 4** Eleven extracted relations with disconnected gene nodes from GeneMANIA

Gene1	Gene2	Confidence	Disconnected node	PMC Excerpt	PMCID/Year	Remark
<i>DCDC1</i>	<i>PAX6</i>	0.96	<i>DCDC1</i>	ELP4 and <i>DCDC1</i> are located downstream of <i>PAX6</i> .	PMC2375324/2008	
<i>ALB</i>	<i>ELP4</i>	0.93	<i>ELP4</i>	<i>ALB</i> was used to normalize <i>ELP4</i> and <i>PAX6</i> values for the detection of the relative copy number of the deletion region.	PMC3859656/2013	
<i>ATOH7</i>	<i>FBN1</i>	0.88	<i>ATOH7</i>	We found 10 candidate POAG genes that were highly expressed in both the CPE and NPE ( <i>AKAP13</i> , <i>C1QBP</i> , <i>CHSY1</i> , <i>COL8A2</i> , <i>CYP1B1</i> , <i>FBN1</i> , <i>IBTK</i> , <i>MFN2</i> , <i>TMCO1</i> , and <i>TMEM248</i> ), three genes that were expressed significantly higher in the CPE ( <i>CDH1</i> , <i>CDKN2B</i> , and <i>SIX1</i> ), and six genes that were expressed significantly higher in the NPE ( <i>ATOH7</i> , <i>CYP1B1</i> , <i>FBN1</i> , <i>MYOC</i> , <i>PAX6</i> , and <i>SIX6</i> ).	PMC3909915/2014	
<i>FBN1</i>	<i>TMEM248</i>	0.88	<i>TMEM248</i>	We found 10 candidate POAG genes that were highly expressed in both the CPE and NPE ( <i>AKAP13</i> , <i>C1QBP</i> , <i>CHSY1</i> , <i>COL8A2</i> , <i>CYP1B1</i> , <i>FBN1</i> , <i>IBTK</i> , <i>MFN2</i> , <i>TMCO1</i> , and <i>TMEM248</i> ), three genes that were expressed significantly higher in the CPE ( <i>CDH1</i> , <i>CDKN2B</i> , and <i>SIX1</i> ), and six genes that were expressed significantly higher in the NPE ( <i>ATOH7</i> , <i>CYP1B1</i> , <i>FBN1</i> , <i>MYOC</i> , <i>PAX6</i> , and <i>SIX6</i> ).	PMC3909915/2014	
<i>GSK3B</i>	<i>MTHFR</i>	0.85	<i>MTHFR</i>	For example, <i>GSK3B</i> has a direct connection with <i>IL4</i> and a secondary connection with <i>MTHFR</i> .	PMC2653647/2009	
<i>GAPDH</i>	<i>VSX1</i>	0.85	<i>VSX1</i>	Each bar represents the relative expression of <i>VSX1</i> normalized to <i>GAPDH</i> in a different tissue/age; mean $\pm$ SD (Sc: sclera, Co: cornea, Ir: iris, CB: ciliary body, Len: lens, Cho:	PMC2267740/2008	
<i>GLS2</i>	<i>HMGB1</i>	0.80	<i>GLS2</i>	the <i>HMGB1</i> inhibitor <i>GA</i> attenuated diabetes-induced upregulation of <i>HMGB1</i> and downregulation of <i>BDNF</i>	PMC3671668/2013	<i>GLS2</i> is a synonym of <i>GA</i>
<i>SHH</i>	<i>ATOH7</i>	0.78	<i>ATOH7</i>	Thus the <i>SHH</i> and <i>GDF11</i> regulate <i>ATOH7</i> , which in turn regulates <i>Brn3b</i> .	PMC2883590/2010	
<b><i>LMX1B</i></b>	<b><i>COL3A1</i></b>	<b>0.45</b>	<b><i>LMX1B</i></b>	Recent immunohistological studies in NPS patients with severe glomerular disease suggest a possible regulation of type III collagen by <i>LMX1B</i> , while the homozygous	PMC2669506/2007	<i>COL3A1</i> is a synonym of <i>Type_III_collagen</i>
<b><i>NPS</i></b>	<b><i>PAX6</i></b>	<b>0.05</b>	<b><i>NPS</i></b>	Research has demonstrated that retinal neurons and RGCs are mainly comprised of anteriorized NPS that express <i>PAX6</i> and <i>OTX2</i> .	PMC3747054/2013	
<b><i>NPS</i></b>	<b><i>OTX2</i></b>	<b>0.05</b>	<b><i>NPS</i></b>	Research has demonstrated that retinal neurons and RGCs are mainly comprised of anteriorized NPS that express <i>PAX6</i> and <i>OTX2</i>	PMC3747054/2013	

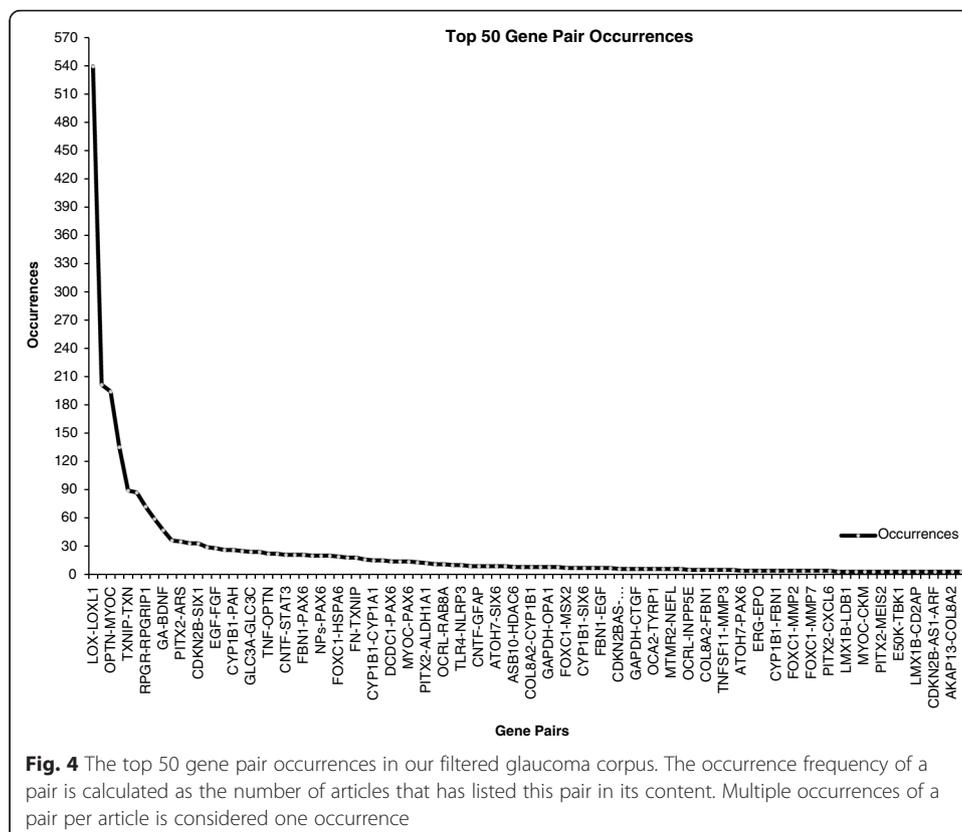
The genes in each extracted relation are listed under the "Gene1" and the "Gene2" columns, respectively. A measure of confidence, reported by ReVerb, is listed under the "Confidence" column and relations with low confidence (<0.5) are bolded. The disconnected node in the relation is listed under the "Disconnected node" column. The associated text that relates the two genes is listed under the "PMC Excerpt" column. Some genes were identified by their synonyms found in either GeneCards or GeneMANIA. The PMCID of the original article, coupled with the year of publication, is given under "PMCID/Year" column. Important remarks and gene synonyms may be listed under the "Remark" column

**Table 5** Percentages of extracted relations

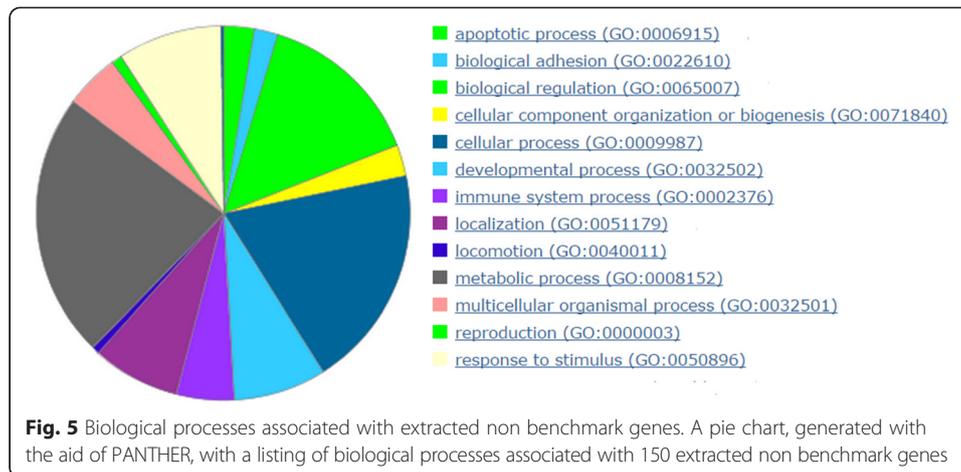
Finding Type	Description	Percentage
Known	Verified	76/257 ~ 30 %
New	Can be verified via one or more indirect paths from the known network	149/257 ~ 58 %
Disconnected	Potential discovery that can be verified by lab experiment in the future	11/257 ~ 4 %
Unverified	Gene symbols could not be found in GeneMANIA, HUGO or GeneCards	21/257 ~ 8 %

The Total number of unique and valid relations is 257, which are classified into known, new, disconnected, and unverified relations, respectively. Description and percentage of each class is given under the "Description" and "Percentage" columns

DAVID, and GeneCodis databases. We excluded the 74 BG from the functional analysis step to avoid intentionally enriching the results with biological processes and pathways that are already known to be related to glaucoma. PANTHER ranked apoptosis at the top of all biological processes associated with those genes (Fig. 5), which is in line with the evidence that retinal ganglion cell death is a hallmark of glaucoma [55]. The most enriched biological processes, associated false discovery rate (FDRs) and enrichment scores, reported by PANTHER and DAVID clustering, are listed in Table 6. Furthermore, PANTHER identified gonadotropin-releasing hormone receptor (GnRHR) (involving 8.1 % of the total genes on average) and Wnt signalling pathways (involving 4.5 % of the total genes on average) with the highest gene associations. Interestingly, it was recently reported that several Wnt signaling target genes have been identified as potential players in glaucoma pathogenesis [56, 57]. The GnRHR pathway was proposed to control central nervous physiology and pathophysiology modulating cognitive changes associated with



**Fig. 4** The top 50 gene pair occurrences in our filtered glaucoma corpus. The occurrence frequency of a pair is calculated as the number of articles that has listed this pair in its content. Multiple occurrences of a pair per article is considered one occurrence



aging and age-related neurodegenerative disorders [58]. Combined pathway analysis by PANTHER and GeneCodis is shown with supporting literature (Fig. 6 and Table 7).

Our result is expected to be comprehensive, with partial resemblance to other studies of glaucoma interaction networks. For example, our result shares only 5 and 29 genes with two previous studies [28, 29] respectively. This emphasizes the fact that interaction networks from text mining approaches can be quite comprehensive because they can incorporate and integrate information from all types of studies. Our enrichment analysis also agreed with previously reported enrichments to glaucoma studies [29] such as apoptosis and induction of apoptosis as underlying biological processes and pathways such as *PDGF* signaling pathway, Ras pathway, and apoptosis signaling pathway.

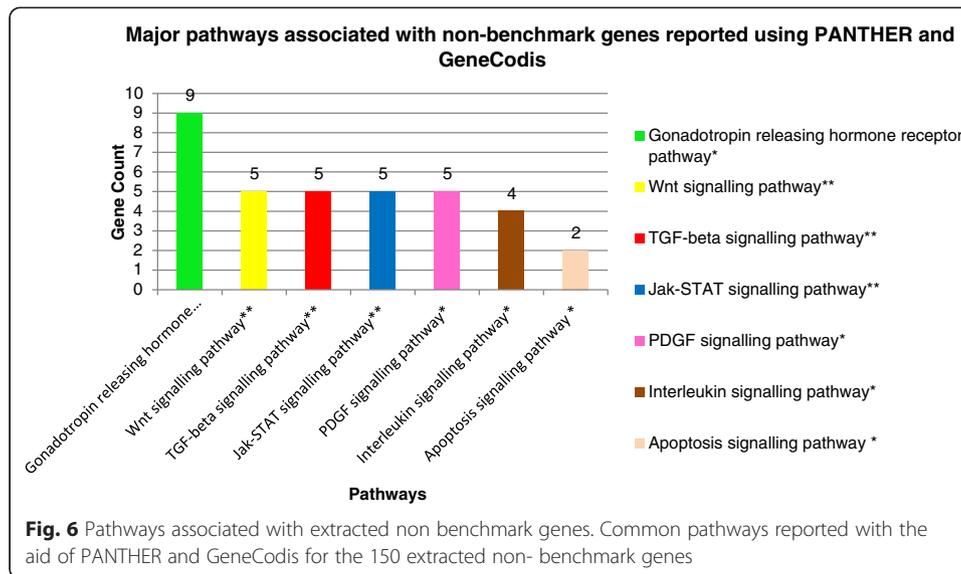
### Network features

The resulting graph is a scale-free network that follows the Barabási–Albert (BA) network model [59]. A scale-free network is a network with node links that follow a power law distribution, i.e. the probability of linking to a given node is proportional to the number of existing links,  $k$ , that node has. Our glaucoma network (Fig. 7) consists of 224 nodes and 255 edges. Network analysis shows that the network has a diameter of 13 and a path length distribution as shown in Fig. 8. While the diameter of the network and path length distribution are quantitative measures that offer insight into how well

**Table 6** Functional analysis of the 150 extracted non-benchmark genes

Biological Process	Gene Count	Corrected <i>P</i> -value
Regulation of apoptosis**	25	2.73E-06
Inflammatory Response*	12	0.002
Immune Response**	17	0.004
Regulation of response to stimulus**	9	0.01
Defense Response*	15	0.01

Biological processes, reported by DAVID, are suffixed by \* and are associated with their genes count and corrected *p*-value. Biological processes, that are common to both PANTHER and DAVID are suffixed by \*\* and are associated with their gene count and corrected *p*-values, obtained from DAVID

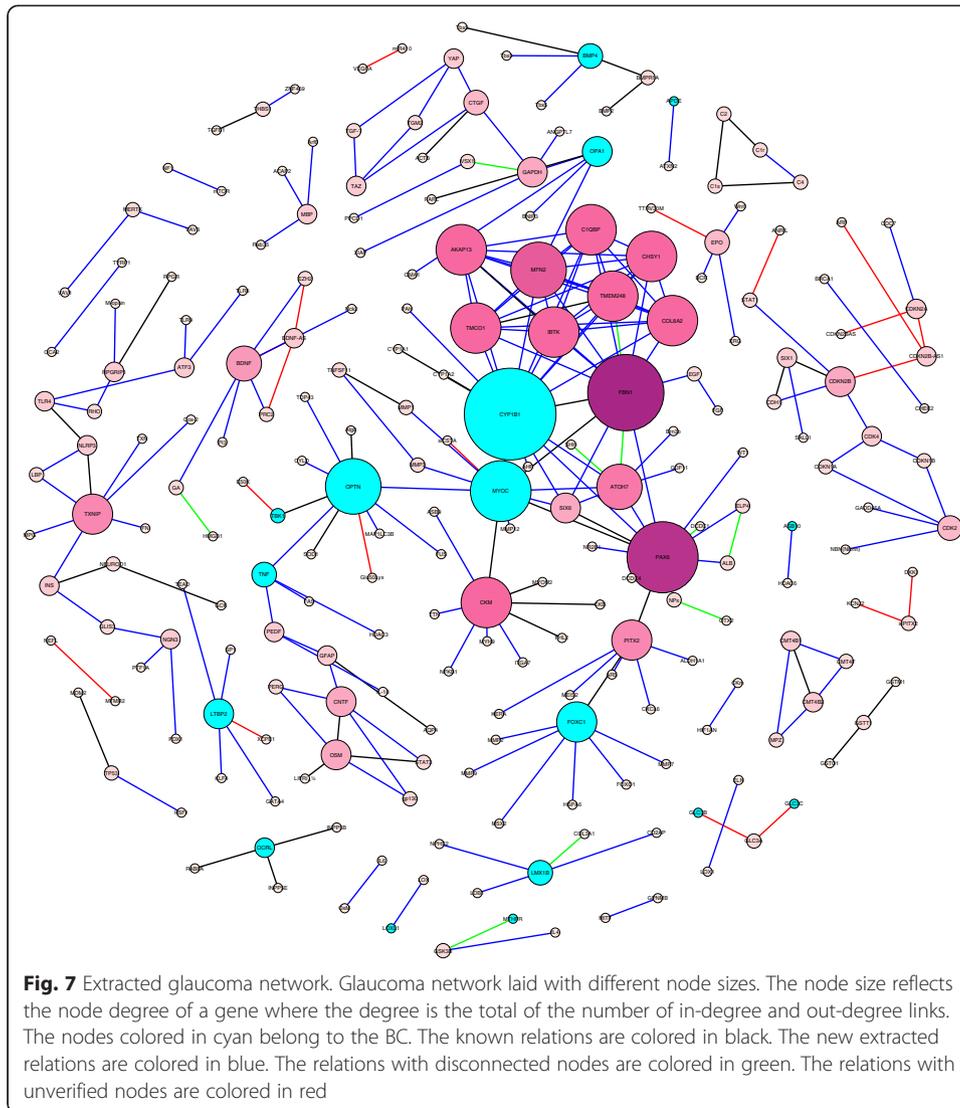


connected a network is, the clustering coefficient describes how clustered the network is. The network diameter is the longest path between all possible pairs of nodes in the network, while the path length distribution summarizes the number of steps along the paths connecting all possible pairs of network nodes. The network has a relatively low clustering coefficient of 0.11; a property which appears to characterize most metabolic networks and protein interaction networks [60, 61], indicating that low degree nodes tend to belong to highly connected neighborhoods, whereas high degree nodes tend to have neighbors that are less connected to each other. The node degree is the number of in-links and out-links for a particular node in the network. The network node degree distribution follows a power law (Fig. 9), another property of scale free networks. Table 8 lists the nodes with top ten degrees, indicating hub entities in the network. To conclude, the current version of the extracted glaucoma interaction network is small but informative. Future versions of the network are expected to evolve closer to a small world network as more links between nodes get added.

**Table 7** Pathway analysis of the 150 extracted NBG

Pathway name	Count of genes in pathway	FDR	% of genes in pathway	Supporting References
Gonadotropin releasing hormone receptor pathway <sup>b</sup>	9		8.1	[58]
Interleukin signaling pathway <sup>b</sup>	6		5.4	[69]
Wnt signalling pathway <sup>a</sup>	5	0.006	4.2	[56, 57]
Jak-STAT signaling pathway <sup>a</sup>	5	0.001	1.8	[70]
PDGF signaling pathway <sup>b</sup>	5		4.5	[71]
TGF-beta signaling pathway <sup>a</sup>	4	0.01	3.6	[72]
Apoptosis signaling pathway <sup>b</sup>	2		1.8	[73, 74]

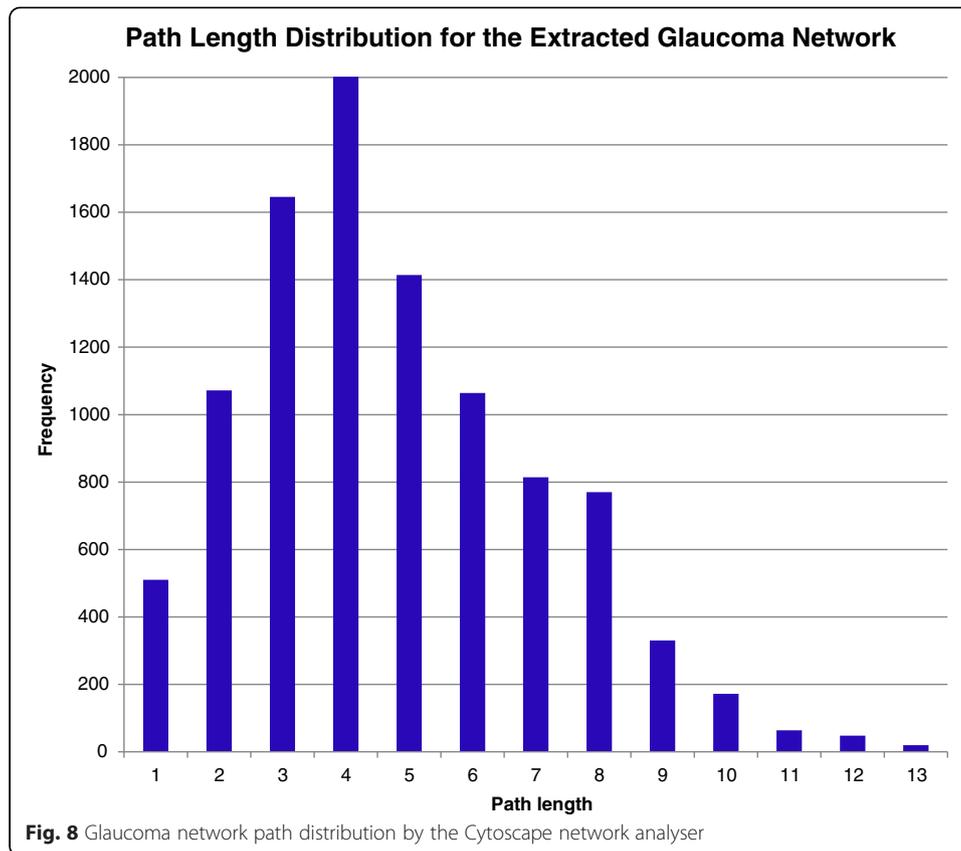
Common pathways, reported by both GeneCodis and PANTHER, are suffixed by <sup>a</sup> and the associated false discovery rate (FDR) from GeneCodis is reported. Pathways, reported by PANTHER, are suffixed by <sup>b</sup>. The percentage of total genes in the pathway is reported along supporting references that link glaucoma to the pathway



### Performance evaluation

As described in the “Methods” section, our text mining pipeline consists of three steps: 1) Text retrieval, 2) Entity extraction, and 3) Relation extraction; each of which has a different associated level of performance. Text retrieval performance is evaluated based on the retrieval of relevant documents. Entity recognition performance is evaluated by the fact that most, if not all genes, should be captured from the collection of glaucoma documents. Relation extraction performance is validated by the extraction of relevant relations. Performance evaluation is usually based on precision (P), recall (R) and F1-score metrics. P is defined as the proportion of retrieved instances that are relevant, while R is the proportion of relevant instances that were retrieved. F1-score combines recall and precision. These metrics are given in Eq (1):

$$P = \frac{\# \text{ of relevant retrieved instances}}{\# \text{ of retrieved instances}}, R = \frac{\# \text{ of relevant retrieved instances}}{\# \text{ of relevant instances}}, F1 = \frac{2 * P * R}{P + R} \quad (1)$$

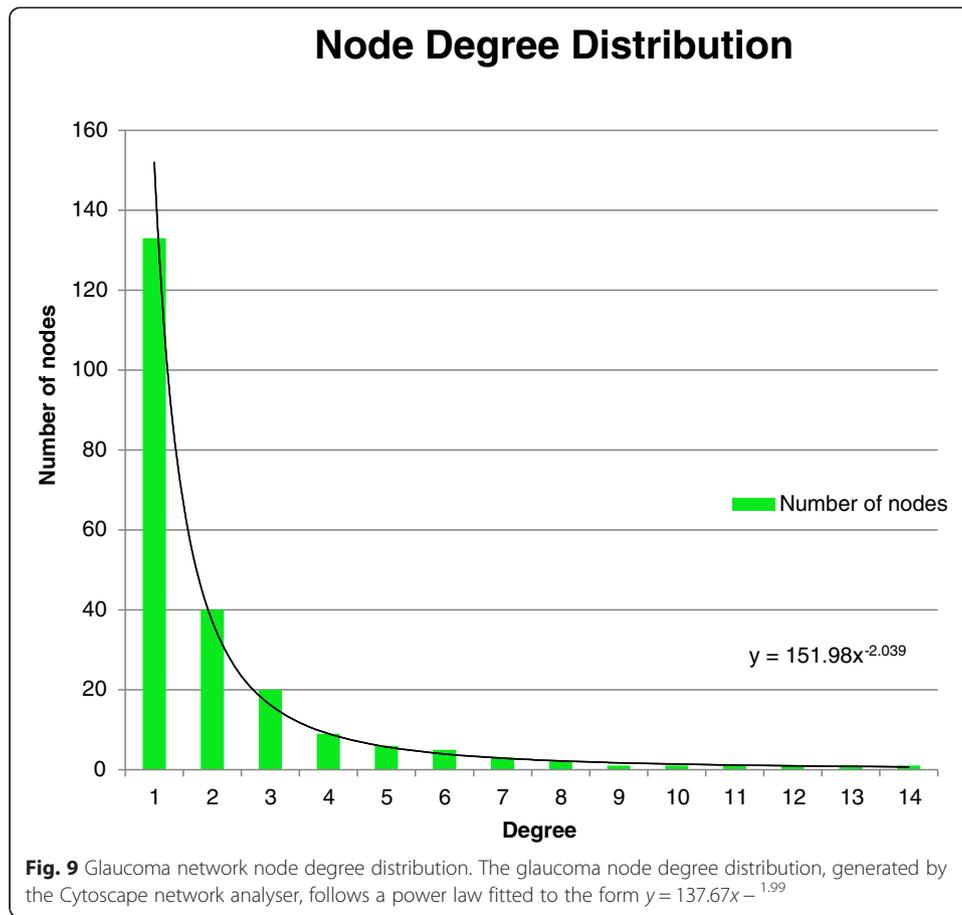


The text retrieval step performance metrics and values are listed in Table 9 and Table 10. For the entity extraction step performance, the GENIA tagger targets a broader domain. Hence, it can be expected to tag varied entities (including localization, cell type, DNA, etc.), but possibly less genes/proteins than the GenTag tagger. This is because the latter is more focused towards genes and proteins. Indeed, in our particular study, GENIA tagger tagged 2410 genes while GenTag tagged 3422 genes. Table 11 lists the performance measures, reported in [62] for GENIA and the average performance measures, reported in [63] and [64] for GenTag.

Because the relation extraction step depends on ReVerb, we report ReVerb's performance from [43], which were 65 % precision and 52 % recall. Therefore, the F1 score associated with the relation extraction step is estimated at 58 %.

## Discussion

While we have described an expansion of the known network of glaucoma related genes, we were surprised that less than a quarter of the genes extracted from DisGeNet and OMIM combined were connected to our network at this time ( $74/305 = 24\%$  BG). Community detection with the Gephi's Louvain modularity maximization algorithm [65], partitioned the network into five distinct modular clusters (Fig. 10). The Louvain modularity maximization algorithm measures the density of links, inside clusters as compared to links between clusters and uses a resolution measure [66] that measures the flows of probabilities in the network. The resulting five clusters formed a strongly connected subnetwork that is 41 % of the size of the original network (96 nodes and



148 edges), with only the giant influential components (nodes with high connectivity) of the network. Examination of the clusters, showed that each has one or more of the BC genes, making a total of 7 BC. Almost the same ratio is observed with the clusters, where less than a quarter of the 30, genes present in both of OMIM and DisGeNET databases ( $7/30 = 23\%$ ), are connected to the clusters. As to the BC genes, the green cluster has *CYP1B1* and *MYOC*, the purple cluster has *OPTN*, *TBK1*, and *TNF*, the red,

**Table 8** Genes (nodes) with the top 10° in the extracted glaucoma interaction network

Gene(node)	Degree
<i>CYP1B1</i>	17
<i>FBN1</i>	14
<i>PAX6</i>	13
<i>MYOC</i>	11
<i>MFN2</i>	10
<i>OPTN</i>	9
<i>CKM</i>	9
<i>AKAP13</i>	9
<i>IBTK</i>	9
<i>TMCO1</i>	9

The degree column represents the total number of a node’s ingoing and outgoing links. Note that *CYP1B1* heads the list with a total of 17 links

**Table 9** Distribution of articles in the text retrieval step, depending on their accessibility and relevance

	Relevant	Not Relevant	Total
Retrieved open access articles	7425	1235	8660
Restricted access (not Retrieved)	22733	unknown	—
Total	31393	—	—

Relevant articles are those that contain at least one occurrence of the word “glaucoma” in their text. The portion of restricted access articles, are not relevant, is unknown to us at the time of writing this article

yellow, and blue clusters have *OPAI*, *FOXC1*, and *CMK* respectively. Their representation here supports the notion that the 30 BC are most highly ranked among all of the BG. Table 12 profiles the different properties of each of the five clusters and Fig. 11 depicts the clusters and their sizes.

The text mining approach, adopted in this study, relies heavily on natural language processing (NLP) methods. We reported in this study, the first version of a glaucoma interaction network, with the intention to report future refined versions when improvements in the text mining pipeline become available. For example, more specificity could likely be added to the results if a better tailored tagger was used. We relied on taggers that were trained on general biological texts that are not specific to glaucoma. Therefore, it is expected that not all entities will be captured from our article collection and an in-house developed tagger, that is trained on literature related to eye diseases and disorders, would likely improve our outcome. Additionally, we note that the currently available glaucoma corpus has a relatively small size compared to other corpora associated with other diseases such as prostate cancer or breast cancer. Since the number of extracted relations is proportional to the size of the corpus, it is desirable to increase the corpus size to discover more relations. There are many possibilities to increase the size of the available glaucoma corpus. For example, PubMed abstracts could be added to the current corpus, or only PubMed abstracts could be considered instead of PMC full text articles. Both options may significantly impact our future results.

Perhaps, the most sought improvement after enlarging the body of literature, would be to reconsider the relation extraction step. ReVerb is designed for open relation extraction, and has not been tweaked for closed relation extraction. In closed relation extraction, the target includes verbs that are known a priori. However, considering our small corpus, it would have negatively affected our extracted relations if we had been confined to a closed set of predetermined verbs [67]. Another difficulty faced by ReVerb is handling complex sentence structures. Although many authors tend to use simple sentence structure such as: Subject-verb-Object, in describing a relationship between two genes, it is not rare for authors to use more complex sentence structures such as conjunctive structure sentences. The latter are sentences that bear multiple verb based relationships or a single verb, to describe many-to-one or one-to-many

**Table 10** Evaluation metrics for the retrieval step

Metric	Value
Precision	$7425/8660 = 85\%$
Recall	$7425/31393 = 23\%$
F1	36%

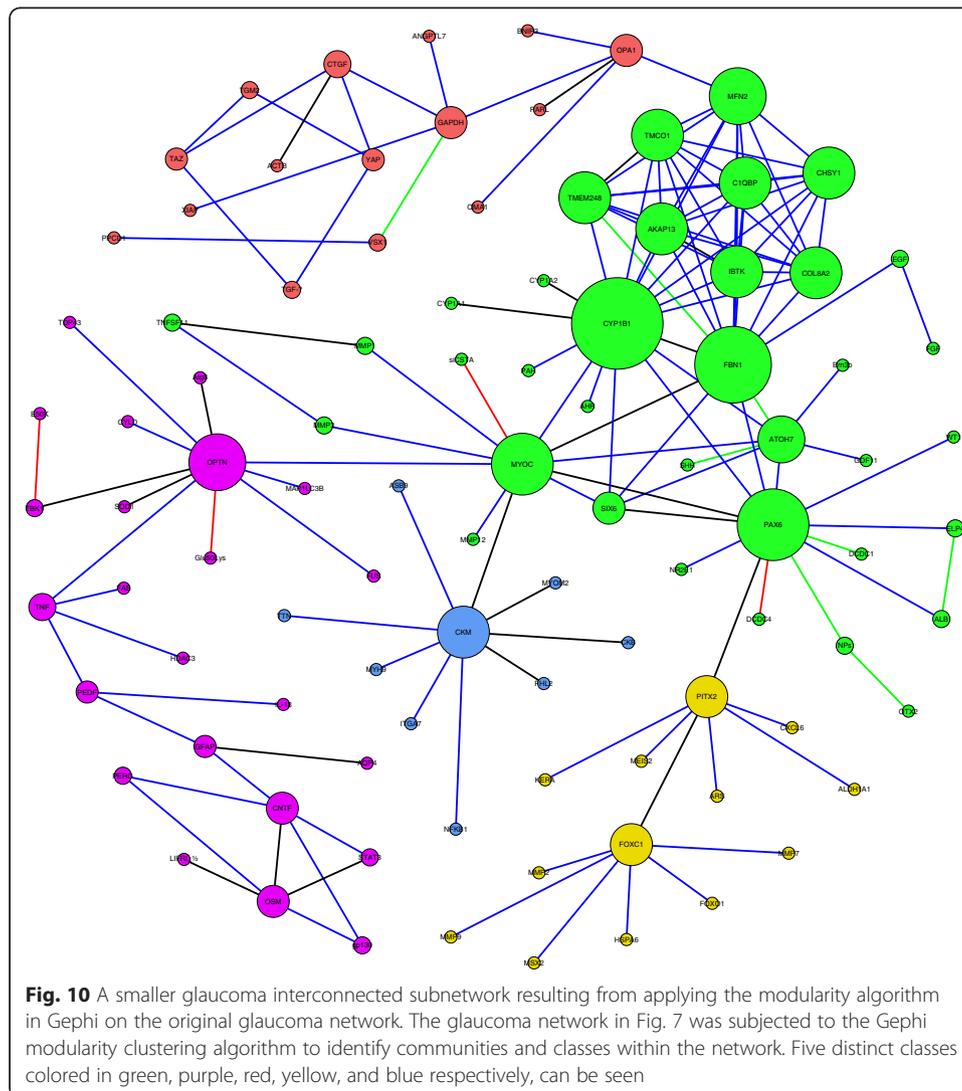
Evaluation metrics are computed based on Table 9. Note that recall is limited by the number of open access articles at this time

**Table 11** Performance measures of the used LingPipe NER tagger

Tagger	Entity Type	Recall (%)	Precision (%)	F-score (%)
GENIA	Protein	81.41	65.82	72.79
	DNA	66.76	65.64	66.2
	RNA	68.64	60.45	64.29
	Cell Line	59.6	56.12	57.81
	Cell Type	70.54	78.51	74.31
	Overall	75.78	67.45	71.37
GENTAG	Gene/Protein	79	88	70.8

Reported measures for the GENIA tagger is based on the GENIA performance web site [62] while performance measures of the GENTAG tagger is the average of the measures reported in [63, 64]

relationships in a single sentence, respectively. Due to its shallow syntactic analysis, Re-Verb’s maximum recall is limited and therefore, it misses most of the conjunctive structure sentences. A better but probably time consuming alternative, is to use an NLP parser such as the Stanford parser [68] to parse target sentences, then search the parsing tree to capture all missing models of verbs.



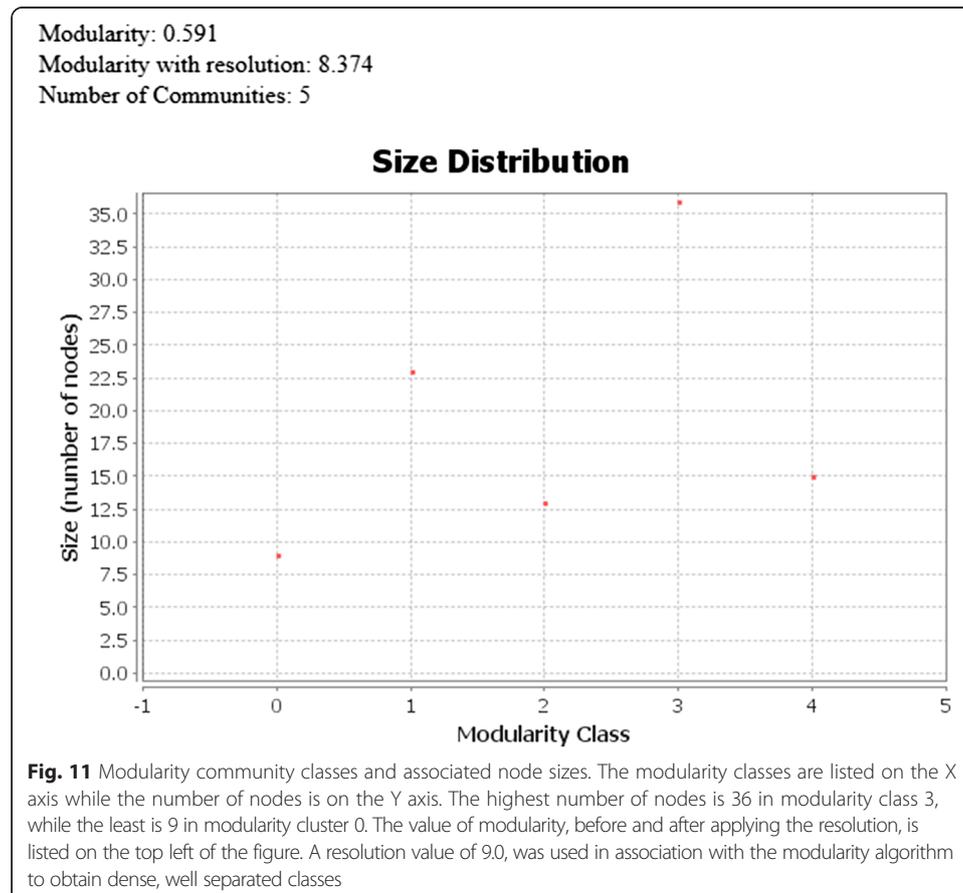
**Table 12** Clusters extracted from the giant components in the glaucoma network and their associated profiles

Cluster	# Nodes	BG	NBG	Node with highest degree	Known relations	New relations	Unverified relations	Disconnected relations
Green	36	6	11	<i>CYP1B1</i> = 17	10	14	8	4
Purple	23	1	9	<i>OPTN</i> = 10	7	15	0	1
Red	15	0	5	<i>OPA1</i> = 5	2	12	0	1
Yellow	13	2	5	<i>FOXC4</i> = 7	2	11	0	0
Blue	9	2	7	<i>CKM</i> = 9	4	5	0	0

The giant components in the glaucoma network depicted in Fig. 7 are clustered into five clusters. Clusters are ordered in descending order of the number of nodes in each cluster. Cluster properties include number of BG, NBG, highest degree, and the number of different types of relations contained within the cluster

## Conclusions

In this study, we have constructed a glaucoma interaction network using a text mining approach applied to open access PMC based literature. Our findings revealed 149 potential new relations. These newly discovered relationships link 74 benchmark genes (BG) present in the 2 databases, DisGeNet and OMIM, with 150 non-benchmark genes (NBG) present in the PubMed Central database, in the form of a small world interaction network. These findings include 21 unverified relations and 11 disconnected relations, which could be verified in the lab. The constructed network contains five distinct gene clusters in association with 7 BC. The 5 clusters are interconnected through 4 gene-gene associations which include: *OPA1-MFN2*, *PITX2-PAX6*, *MYOC-*



*CKM* and *MYOC-OPTN*. Thus the larger network is only possible because of these 4 bridges. It is important to note that 2 of these 4 gene-gene bridges, *OPA1-MFN2* and *MYOC-OPTN*, were discovered through this text mining approach which has associated genes in the DisGeNet and OMIM databases with the PubMed Central database. Finally, we have discussed several important issues with text mining approaches which could aid future iterations of disease-based gene-interaction networks.

### Additional files

**Additional file 1:** Filtered Extracted Relations. (XLSX 58 kb)

**Additional file 2:** Unique Extracted Relations. (XLSX 32 kb)

**Additional file 3:** Glaucoma Benchmark Genes (XLSX 38 kb)

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

MS initiated, designed and implemented the study and drafted the manuscript. ON oversaw the text mining approach and the result validation, and revised the manuscript. NGFC coordinated the study, provided biological interpretation and revisions to the manuscript drafts. All authors read and approved the final manuscript.

### Acknowledgements

We would like to thank Dr. Eric Rouchka for his valuable comments for improving the manuscript. This work was supported in part by grants from the National Eye Institute R01EY017594 and the National Institute of General Medical Sciences P20 GM103436.

### Author details

<sup>1</sup>Department of Anatomical Sciences and Neurobiology, University of Louisville, School of Medicine, Louisville, KY, USA.

<sup>2</sup>Knowledge Discovery & Web Mining Lab, Department of Computer Engineering & Computer Science, University of Louisville, J.B Speed School of Engineering, Louisville, KY, USA.

Received: 9 October 2015 Accepted: 23 April 2016

Published online: 05 May 2016

### References

- Christopher R, Dhiman A, Fox J, Gendelman R, Haberichter T, Kagle D, Spizz G, Khalil IG, Hill C. Data-driven computer simulation of human cancer cell. *Ann N Y Acad Sci.* 2004;1020:132–53.
- Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med.* 1986;30(1):7–18.
- Srinivasan P, Libbus B. Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics.* 2004;20 Suppl 1:i290–296.
- Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR. Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics.* 2004;20(3):389–98.
- Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics.* 2004;5:147.
- van der Eijk CC, van Mulligen EM, Kors JA, Mons B, van den Berg J. Constructing an associative concept space for literature-based discovery. *J Am Society Information Science Technology.* 2004;55(5):436–44.
- Zaremba S, Ramos-Santacruz M, Hampton T, Shetty P, Fedorko J, Whitmore J, Greene JM, Perna NT, Glasner JD, Plunkett 3rd G, et al. Text-mining of PubMed abstracts by natural language processing to create a public knowledge base on molecular mechanisms of bacterial enteropathogens. *BMC Bioinformatics.* 2009;10:177.
- Abulaish M, Dey L. Biological relation extraction and query answering from medline abstracts using ontology-based text mining. *Data Knowledge Engineering.* 2007;61(2):228–62.
- He M, Wang Y, Li W. PPI finder: a mining tool for human protein-protein interactions. *PLoS One.* 2009;4(2):e4554.
- Tudor CO, Ross KE, Li G, Vijay-Shanker K, Wu CH, Arighi CN. Construction of phosphorylation interaction networks by text mining of full-length articles using the eFIP system. *Database.* 2015;2015:bav020.
- Yang Y, Wang Y, Zhou K, Hong A. Constructing regulatory networks to identify biomarkers for insulin resistance. *Gene.* 2014;539(1):68–74.
- Malhotra A, Younesi E, Bagewadi S, Hofmann-Apitius M. Linking hypothetical knowledge patterns to disease molecular signatures for biomarker discovery in Alzheimer's disease. *Genome Med.* 2014;6(11):97.
- Quan C, Ren F. Gene-disease association extraction by text mining and network analysis. In: *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EAACL.* 2014. p. 54–63.
- Ozgur A, Vu T, Erkan G, Radev DR. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics.* 2008;24(13):i277–285.
- Wu X, Chen L, Wang X. Network biomarkers, interaction networks and dynamical network biomarkers in respiratory diseases. *Clin Transl Med.* 2014;3:16.
- Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, 23 Rosse C. Relations in biomedical ontologies. *Genome Biol.* 2005;6(5):R46.

17. Skusa A, Rüegg A, Köhler J. Extraction of biological interaction networks from scientific literature. *Brief Bioinform.* 2005;6(3):263–76.
18. Nguyen N, Miwa M, Tsuruoka Y, Tojo S. Open information extraction from biomedical literature using predicate-argument structure patterns. In: *Proceedings of The 5th International Symposium on Languages in Biology and Medicine*. 2013. p. 51–5.
19. Etzioni O, Banko M, Soderland S, Weld DS. Open information extraction from the web. *Communications ACM.* 2008;51(12):68–74.
20. Rinaldi F, Clemenide S, Marques H, Ellendorff T, Romacker M, Rodriguez-Esteban R. OntoGene web services for biomedical text mining. *BMC Bioinformatics.* 2014;15(14):S6.
21. Jelier R, Schuemie MJ, Veldhoven A, Dorssers LC, Jenster G, Kors JA. Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biol.* 2008;9(6):R96.
22. Torii M, Li G, Li Z, Oughtred R, Diella F, Celen I, Arighi CN, Huang H, Vijay-Shanker K, Wu CH. RLIMS-P: an online text-mining tool for literature-based extraction of protein phosphorylation information. *Database.* 2014;2014:bau081.
23. Guo Y, Séaghdha DO, Silins I, Sun L, Högberg J, Stenius U, Korhonen A. CRAB 2.0: A text mining tool for supporting literature review in chemical cancer risk assessment. *COLING.* 2014;2014:76.
24. Kingman S. Glaucoma is second leading cause of blindness globally. *Bull World Health Organ.* 2004;82(11):887–8.
25. Beidoe G, Mousa SA. Current primary open-angle glaucoma treatments and future directions. *Clin Ophthalmol.* 2012;6:1699–707.
26. HU T, Darabos C, Cricco Me KE, Moore JH. Genome-wide genetic interaction analysis of glaucoma using expert knowledge derived from human phenotype networks. In: *Pacific Symposium on Biocomputing Pacific Symposium on Bioinformatics.* 2014. p. 207–18. World Scientific.
27. Basu K, Sen A, Ray K, Ghosh I, Datta K, Mukhopadhyay A. Genetic association and gene-gene interaction of HAS2, HABP1 and HYAL3 implicate hyaluronan metabolic genes in glaucomatous neurodegeneration. *Dis Markers.* 2012; 33(3):145–54.
28. Colak D, Morales J, Bosley TM, Al-Bakheet A, AlYounes B, Kaya N, Abu-Amero KK. Genome-Wide Expression Profiling of Patients with Primary Open Angle GlaucomaGene Expression Profiling of POAG. *Invest Ophthalmol Vis Sci.* 2012;53(9):5899–904.
29. Nikolskaya T, Nikolsky Y, Serebryskaya T, Zvereva S, Sviridov E, Dezso Z, Rahkmatulin E, Brennan RJ, Yankovsky N, Bhattacharya SK. Network analysis of human glaucomatous optic nerve head astrocytes. *BMC Med Genomics.* 2009;2(1):24.
30. Ronen F, James S. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured*. New York, NY, USA: Cambridge University Press; 2006.
31. Mooney RJ, Bunescu R. Mining knowledge from text using information extraction. *ACM SIGKDD Explorations Newsletter.* 2005;7(1):3–10.
32. The PMC Open Access Subset [<http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>]. Accessed 25 Mar 2015.
33. Pyysalo S, Ohta T, Tsujii J. An analysis of gene/protein associations at PubMed scale. *J Biomed Semantics.* 2011; 2(5):S5.
34. Baldwin B, Carpenter B. LingPipe. 2003. Available from World Wide Web: <http://alias-i.com/lingpipe/>. Accessed 25 Mar 2015.
35. Tanabe L, Xie N, Thom LH, Matten W, Wilbur WJ. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics.* 2005;6(1):S3.
36. Kim JD, Ohta T, Tsujii J. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics.* 2008;9:10.
37. Krallinger M, Leitner F, Valencia A. Assessment of the second BioCreative PPI task: automatic extraction of protein-protein interactions. In: *Proceedings of the second biocreative challenge evaluation workshop*. 2007. p. 41–54.
38. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 2005;33 suppl 1:D514–7.
39. Pinero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F, Furlong LI. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database (Oxford).* 2015;bav028.
40. Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F, Furlong LI. Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PLoS One.* 2011;6(6):e20284.
41. Bauer-Mehren A, Rautschka M, Sanz F, Furlong LI. DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Bioinformatics.* 2010;26(22):2924–6.
42. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Research.* 2015;43(D1):D1079–85.
43. Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2011. p. 1535–45. Association for Computational Linguistics.
44. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 2010;38(Web Server issue):W214–220.
45. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 2006;34 suppl 1:D535–9.
46. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. *Trends Genet.* 1997;13(4):163.
47. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. *ICWSM.* 2009;8:361–2.
48. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11): 2498–504.

49. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc.* 2013;8(8):1551–66.
50. da Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 2009;4(1):44–57.
51. da Huang W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 2009;37(1):1–13.
52. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol.* 2007;8(1):R3.
53. Nogales-Cadenas R, Carmona-Saez P, Vazquez M, Vicente C, Yang X, Tirado F, Carazo JM, Pascual-Montano A. GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Res.* 2009;37 suppl 2:W317–22.
54. Tabas-Madrid D, Nogales-Cadenas R, Pascual-Montano A. GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic Acids Res.* 2012;40(W1):W478–83.
55. Rokicki W, Dorecka M, Romaniuk W. Retinal ganglion cells death in glaucoma—mechanism and potential treatment. Part I. *Klin Oczna.* 2006;109(7–9):349–52.
56. Wang WH, McNatt LG, Pang IH, Millar JC, Hellberg PE, Hellberg MH, Steely HT, Rubin JS, Fingert JH, Sheffield VC, et al. Increased expression of the WNT antagonist sFRP-1 in glaucoma elevates intraocular pressure. *J Clin Invest.* 2008; 118(3):1056–64.
57. Villarreal Jr G, Chatterjee A, Oh SS, Oh DJ, Kang MH, Rhee DJ. Canonical wnt signaling regulates extracellular matrix expression in the trabecular meshwork. *Invest Ophthalmol Vis Sci.* 2014;55(11):7433–40.
58. Wang L, Chadwick W, Park SS, Zhou Y, Silver N, Martin B, Maudsley S. Gonadotropin-releasing hormone receptor system: modulatory role in aging and neurodegeneration. *CNS Neurol Disord Drug Targets.* 2010;9(5):651–60.
59. Barabási A-L, Albert R. Emergence of scaling in random networks. *Science.* 1999;286(5439):509–12.
60. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L. Hierarchical organization of modularity in metabolic networks. *Science.* 2002;297(5586):1551–5.
61. Yook SH, Oltvai ZN, Barabási AL. Functional and topological characterization of protein interaction networks. *Proteomics.* 2004;4(4):928–42.
62. GENIA Tagger- part-of-speech tagging, shallow parsing, and named entity recognition for biomedical text- [<http://www.nactem.ac.uk/tsujii/GENIA/tagger/>]. Accessed 25 Mar 2015.
63. Chtioui S. Evaluation of gene/protein name recognition Programs. Geneva: Masters in Proteomics and Bioinformatics, University of Geneva; 2008.
64. Ekbal A, Saha S, Sikdar UK. Biomedical named entity extraction: some issues of corpus compatibilities. *Springerplus.* 2013;2:601.
65. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Statistical Mechanics.* 2008;2008(10):10008.
66. Lambiotte R, Delvenne JC, Barahona M. Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770.* 2008.
67. Pyysalo S, Ohta T, Kim J-D, Tsujii J. Static relations: a piece in the biomedical information extraction puzzle. In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing.* 2009. p. 1–9. Association for Computational Linguistics.
68. De Marneffe M-C, MacCartney B, Manning CD. Generating typed dependency parses from phrase structure parses. In: *Proceedings of LREC.* 2006. p. 449–54.
69. Nakatake S, Yoshida S, Nakao S, Arita R, Yasuda M, Kita T. Hyphema is a risk factor for failure of trabeculectomy in neovascular glaucoma: a retrospective analysis. *BMC Ophthalmol.* 2014;14(1):55.
70. Wang DY, Ray A, Rodgers K, Ergorul C, Hyman BT, Huang W. Global gene expression changes in rat retinal ganglion cells in experimental glaucoma. *Invest Ophthalmol Vis Sci.* 2010;51(8):4084–95.
71. Stewart MW. PDGF: ophthalmology's next great target. 2013.
72. Wecker T, Han H, Borner J, Grehn F, Schlunck G. Effects of TGF-beta2 on cadherins and beta-catenin in human trabecular meshwork cells. *Invest Ophthalmol Vis Sci.* 2013;54(10):6456–62.
73. Ayub H, Micheal S, Akhtar F, Khan MI, Bashir S, Waheed NK, Ali M, Schoenmaker-Koller FE, Shafique S, Qamar R, den Hollander AI. Association of a Polymorphism in the BIRC6 Gene with Pseudoexfoliative Glaucoma. *PLoS One.* 2014;9(8):e105023.
74. Izzotti A, Longobardi M, Cartiglia C, Sacca SC. Mitochondrial damage in the trabecular meshwork occurs only in primary open-angle glaucoma and in pseudoexfoliative glaucoma. *Plos One.* 2011;6(1):e14567.

Submit your next manuscript to BioMed Central  
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

